Unsupervised Learning

Unsupervised Learning

Clustering

1. Clustering is primarily an exploratory technique to discover hidden structures of the data, possibly as a prelude to more focused analysis or decision processes
    a. A way to decompose a data set into subsets with each subset representing a group with similar characteristics

    b. Group such that objects in the same group are more similar to each other in some sense than to objects of different groups

    c. The groups are known as clusters and each cluster gets distinct label called cluster id, the centroid of the cluster, and other details

2. Clustering is often used as a lead-in to classification. Once the clusters are identified, labels can be applied to each cluster to classify each group based on its characteristics

Applications of clustering

Some specific applications of k-means are image processing, medical, and customer segmentation

a. **Image processing** : used to cluster of pixels representing objects in each frame. The attributes of each pixel can include brightness, color, and location, the x and y coordinates in the frame. Successive frames are examined to identify any changes to the clusters. These newly identified clusters may indicate unauthorized access to a facility.

b. **Medical** : Patient attributes such as age, height, weight, systolic and diastolic blood pressures, cholesterol level, and other attributes can identify naturally occurring clusters under various health conditions

c. **Customer segmentation** : Cluster customers on basis of frequency of purchase, recency of purchase, value of purchase and look for common attributes among high value customers. Target all potential customers who have similar attributes

Clustering types

1. Two broad categories of clustering include hierarchical (agglomerative, divisive) and non hierarchical

2. Hierarchical clustering
    a) Agglomerative clustering algorithm uses a bottom-up approach and merges smaller clusters into larger ones
    b) Divisive clustering uses top-bottom approach to break a large cluster into smaller clusters

3. Non-hierarchical / partitional clusters are formed on assumption that the clusters are disjoint and there is no hierarchical relation between them. K Means is an example

Clustering – Distance calculations

1. Irrespective of the clustering algorithm, we need a way of defining and calculating distance between two data points in mathematical space (records in the database)

2. We also need a way to define and calculate distance between clusters

3. Distance between two data points is a measure of similarity between the points

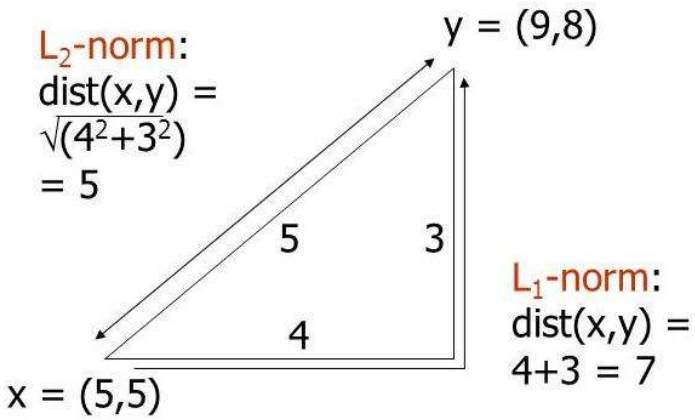
4. The lesser the distance, more similar the data points are

5. There are many ways of calculating distance between two points i.e. if d = f(x,y) then there are many ways in which f can be implemented

Euclidean Distance

1. Euclidean Distance
   a. L2 norm : d(x,y) = square root of the sum of the squares of the differences between x and y in each dimension. The most common notion of “distance”. If there are two dimensions x and y, the distance between two point A and B is –

   b. L1 norm : sum of the differences in each dimension. Manhattan distance = distance if you had to travel along coordinates only



Non Euclidean Distance

2. Non Euclidean Distance
    a. Jaccard distance for sets = 1 minus ratio of sizes of intersection and union.
    b. Cosine distance = angle between vectors from the origin to the points in question.
    c. Edit distance = number of inserts and deletes to change one string into another
    d. Mahalanobis distance

3. Normalizing the numerical measurements
    a. The measures computed in Euclidian methods are highly influenced by the scale of each variable
    b. Variables with larger scale have much greater influence over the total distance
    c. Hence all the measurements are converted to same scale (convert to z scores for e.g.)

Note: distance measurement is just a way of assessing similarity/ dissimilarity. The common parlance of distance will not help in understanding other methods of distance calculations

Distance measures

1. Distance measures and some key points:
    a. Choice of distance measures play a key role in cluster analysis.
    b. Knowledge of the distribution of data (gaussian or otherwise) will help
    c. Are the various attributes independent or influence eachother
    d. Are their outliers in the data on the various dimensions
    e. Though Euclidian distance is the most commonly used distance metric, it has three main features that should be kept in view
        a. It is highly scale dependent. Changing the units of one variable can have a huge influence on the results. Hence standardizing the dimensions is a good practice
        b. It completely ignores the relationship between measurements (Refer to Mahalanobis distance diagram)
        c. It is sensitive to outliers. If the data has outliers that cannot be handled or removed, use of Manhattan distance is preferred

K Means Clustering – Some considerations

1. K-Means (a.k.a Lloyd's algorithm) clusters data by separating data points into groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squared errors

2. It requires the number of clusters to be specified, hence the term "K" in its name

3. It divides the samples into K disjoint clusters Ci, each described by the mean of the samples in the cluster. The means are commonly called "centroids" (they are not the points from the data)

4. The K-Means algorithm chooses centroids that minimizes the inertia across all the clusters

K Means Clustering – Some considerations Pt 2

5. From a computational perspective, the k-means algorithm is indifferent to the units of measure for a given attribute (for example, meters or centimeters for a patient's height). However, the algorithm will identify different clusters depending on the choice of the units of measure.

6. Choosing different starting points can result in different clusters. The algorithm is sensitive to the initial starting condition

7. Given enough time, K-means will always converge, however this may be a local minimum. This is highly dependent on the initialization of the centroids

8. Scikit-learn has implemented K-mean++ initialization scheme, which initializes centroids to be distant to one another which provably leads to better results

K-means Clustering objective

The objective function of clustering is –

- Given: a set of n observations $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, where each observation is a d-dimensional real vector
- Given: a number of clusters k
- Compute: a cluster assignment mapping $C(\boldsymbol{x}_i) \in \{1, \ldots, k\}$ that minimizes the **within cluster sum of squares (WCSS)**:

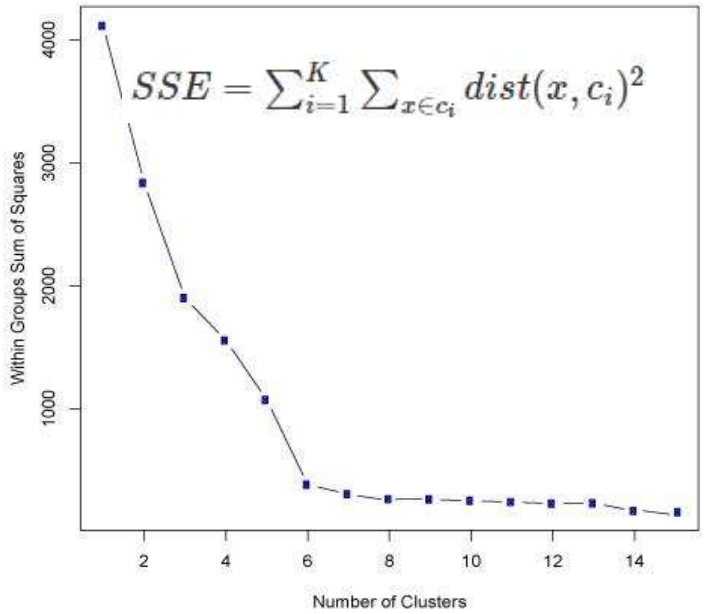
$$\sum_{i=1}^{n} \left\| \boldsymbol{x}_i - \boldsymbol{\mu}_{C(\boldsymbol{x}_i)} \right\|^2$$

where **centroid** $\boldsymbol{\mu}_{C(\boldsymbol{x}_i)}$ is the mean of the points in cluster $C(\boldsymbol{x}_i)$

Elbow method

Elbow Method

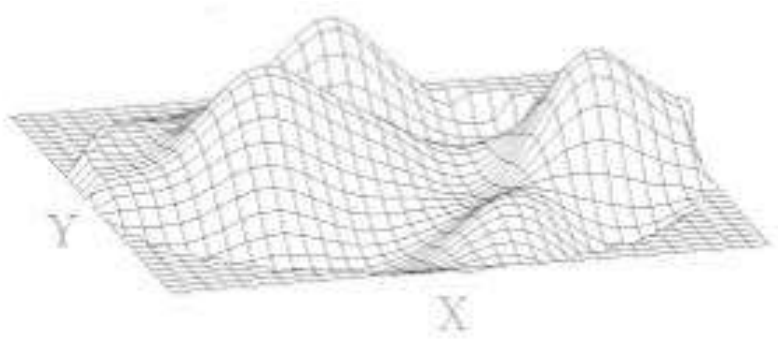
7. Without a priori knowledge, one can use elbow method that measures the homogeneity or heterogeneity within clusters as the number of clusters change (i.e. K is changed). One way to measure is use sum of square errors in each cluster



Visual Analysis for Clustering

Visual Analysis for Clustering

1. Visual analysis of the attributes selected for the clustering may given an idea of the range of values that K should be evaluated in



2. Identifying the attributes on which clusters are clearly demarcated and using them in incremental order to build the multi-dimensional clusters likely to give much better clusters than using all the attributes at one go

Dynamic Clustering

Dynamic Clustering

1. Clustering on correct attributes is the key to good clustering results.

2. We can also consider those attributes who's value changes with time. For e.g. age, income category, years of work experience etc.

3. We can use sequential k means clustering over time to track individual clusters (how they change in size, shape and position

4. We can also understand how individual data points move across clusters, form new clusters etc.

5. Analyzing the changes in the clusters over time using metrics such as

6. Cluster size, new entries and exits one can analyze the impact of strategies designed based on earlier clustering analysis

K-Means Clustering Strengths and Weakness

K-Means Clustering Strengths and Weakness

Strengths	Weakness
Use simple principles without the need for any complex statistical terms	Computationally intensive How to fix K?
Once clusters and their associated centroids are identified, it is easy to assign new objects (for example, new customers) to a cluster based on the object's distance from the closest centroid	The k-means algorithm is sensitive to the starting positions of the initial centroid. Thus, it is important to rerun the k-means analysis several times for a particular value of k to ensure the cluster results provide the overall minimum WSS
Because the method is unsupervised, using k-means helps to eliminate subjectivity from the analysis.	Susceptible to curse of dimensionality